GenoDup Pipeline: a tool to detect genome duplication using the dS-based method

GenoDup Pipeline: a tool to detect genome duplication using the dS-based method

Yafei Mao

Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, Japan

ABSTRACT

Understanding whole genome duplication (WGD), or polyploidy, is fundamental to investigating the origin and diversification of organisms in evolutionary biology. The wealth of genomic data generated by next generation sequencing (NGS) has resulted in an urgent need for handy and accurate tools to detect WGD. Here, I present a useful and user-friendly pipeline called GenoDup for inferring WGD using the dS-based method. I have successfully applied GenoDup to identify WGD in empirical data from both plants and animals. The GenoDup Pipeline provides a reliable and useful tool to infer WGD from NGS data.

Subjects Bioinformatics, Evolutionary Studies, Genomics

Keywords Polyploidy, Whole-genome duplication (WGD), dS-based, Next generation sequencing (NGS), Software, Age distribution, GenoDup

INTRODUCTION

Whole (large-scale)-genome duplication (WGD), or polyploidy, has been regarded as an evolutionary landmark in the origin and diversification of animals, plants, and other evolutionary lineages (*Van De Peer, Mizrachi & Marchal, 2017*). Previous studies have shown that WGD plays an important role in enhancing speciation and reducing risks of extinction. Moreover, evolutionary novelty can be generated by duplicated genes via subfunctionalization, neofunctionalization, and dosage effects under WGD (*Glasauer & Neuhauss, 2014*; *Van De Peer, Mizrachi & Marchal, 2017*). Therefore, identification of WGD is the first step to understanding the impacts of WGD and the fates of duplicated genes. WGD is now known to be a common event in plants, since the availability of genomic data generated by next generation sequencing (NGS) (*Jiao, 2018*; *Jiao & Paterson, 2014*; *Soltis et al., 2009*; *Soltis et al., 2015*). Meanwhile, recent studies also suggest that WGD is a common evolutionary force in animals (*Li et al., 2018*; *Van De Peer, Mizrachi & Marchal, 2017*). Hence, an easy-to-use pipeline is urgently needed to infer WGD using NGS data.

There are three main approaches to infer WGD with NGS data (*Tiley, Ane & Burleigh, 2016*). First, identification of synteny blocks is the most straightforward method to detect WGD, but it requires high-quality genome assembly, and sadly, many genomes have not yet reached that assembly quality (*Bowers et al., 2003*). Second, phylogenetic analysis of gene families can unravel WGD when organisms have undergone extensive gene loss or genome shuffling (*Jiao et al., 2011*; *Jiao et al., 2014*), but the uncertainty of gene tree reconstruction

Submitted 5 October 2018
Accepted 16 December 2018
Published 23 January 2019

Corresponding author
Yafei Mao, yafei.mao@oist.jp

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 8

DOI **10.7717/peerj.6303**

is a serious limitation as well as heavy computation is required. Finally, analysis of rates of synonymous substitutions per synonymous site (dS) of duplicated genes (the dS-based method or age distribution method) is the most common and widely used approach to infer WGD (*Lynch & Conery, 2000*; *Tiley, Barker & Gordon Burleigh, 2018*).

Synonymous substitutions are usually under little selection, thus, rates of synonymous substitutions per synonymous site (dS) between two genes can be regarded as a proxy for the time of their divergence (*Lynch & Conery, 2000*). In addition, the process of gene duplication and loss is assumed as a steady birth and death model. Therefore, the distribution of dS values for paralogs should be an "L-shape" curve (*Maere et al., 2005*). WGD usually generates numerous paralogs simultaneously and thus a peak in the distribution of dS values can be considered as a WGD event. Compared to phylogenomic and synteny block approaches, assembled genome information and heavy computation are not required for the dS-based method. Yet because of the effects of synonymous substitution saturation and gene retention rates, it is difficult for the ds-based method to infer WGD events, which are too ancient or under lower gene retention rates (*Tiley, Barker & Gordon Burleigh, 2018*; *Van De Peer, Mizrachi & Marchal, 2017*). Despite all that, the dS-based method is still a relatively quick and easy way to infer WGD as the first step, and then the inferred WGD could be confirmed by phylogenomic and synteny block approaches.

The dS-based method is a fragmented step-wise process (*Vanneste et al., 2014*). Multiple software packages are required to build gene pairs, align sequences, and calculate dS values. Usually, there are two ways to build gene pairs using the dS-based method. The first one is to use paralogous gene pairs, generated by gene family cluster (orthogroup) information or gene pair information. Gene pair information is usually created by all-against-all BLAST directly while orthogroup information can be generated by a clustering algorithm based on all-against-all BLAST result (e.g., OrthoMCL *Li, Stoeckert Jr & Roos, 2003*, OrthoFinder *Emms & Kelly, 2015*). Generally, orthogroups provide more accurate information of duplicated genes rather than gene pairs. Secondly, paralogs located at the same synteny block are considered as anchor gene pairs. Thus, we could use synteny information to generate anchor gene pairs. Anchor gene pairs accurately represent duplicated genes and usually provide more information about ancient duplication events. Together, consistent results from both approaches yield a credible conclusion for the dS-based method.

DupPipe is a web-based method to infer WGD using the dS-based method (*Barker et al., 2010*). In addition, FASTKs is a pipeline to calculate dS values for gene pairs (*McKain et al., 2016*). Both DupPipe and FASTKs calculate dS values based on gene pair information, but not based on orthogroup information, to infer WGD. Here, an open-source script called GenoDup Pipeline is developed to infer WGD using the dS-based method based on orthogroup or/and (paralogous or anchor) gene pair information.

MATERIALS & METHODS

GenoDup Pipeline architecture

GenoDup Pipeline is written in Python integrating with alignment of sequences, building gene pairs, and dS value calculations. BioPython must be installed and three more executable

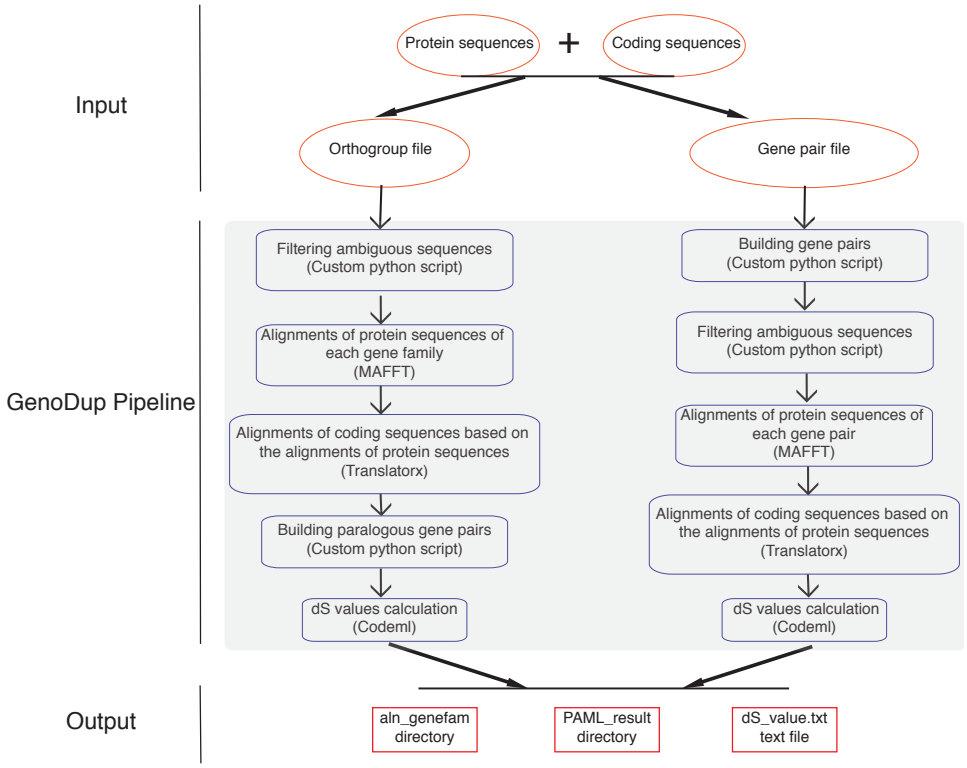

Figure 1 Workflow in GenoDup Pipeline. Orange oval boxes represent inputs. Blue boxes represent three subroutines in the GenoDup Pipeline. Red boxes represent outputs generated by GenoDup Pipeline.
Full-size DOI: 10.7717/peerj.6303/fig-1

dependencies are needed: MAFFT (*Katoh et al., 2002*), Translatorx (*Abascal, Zardoya & Telford, 2010*), and Codeml package in PAML (*Yang, 2007*). Nuclear protein coding sequences (CDS) and the corresponding protein sequences are mandatory inputs to run GenoDup Pipeline. In addition, orthogroup information or gene pair information is another mandatory input for orthogroup and gene pair approach, respectively. Once mandatory files have been inputted appropriately, 3 subroutines in GenoDup run as follows (Fig. 1).

(1) Alignment of gene pairs: GenoDup Pipeline can automatically align gene pairs or gene families using MAFFT and Translatorx. Before performing gene pair or gene family alignments, GenoDup firstly filters ambiguous sequences that contain 'N', and removes CDS that mis-match to the corresponding protein sequences. Then, MAFFT is used to perform alignments of protein sequences with parameters: localpair and maxiterate: 1000; and Translatorx is used to align CDS based on alignments of the corresponding protein sequences.

(2) Building gene pairs: there are two ways to build gene pairs in GenoDup Pipeline. The first way requires an orthogroup information file and a number (N) as inputs. Only the orthogroups, which contain less than N genes, can be used to build gene pairs. GenoDup Pipeline builds n(n-1)/2 paralogous gene pairs within a gene family (n is the number of genes in a gene family). OrthoMCL is recommended to generate orthogroup

(*Li, Stoeckert Jr & Roos, 2003*). The second way requires a gene pair information file. Paralogous gene pairs can be generated by all-against-all BLAST. Or, MCScanX (*Wang et al., 2012*) and i-ADHoRe (*Proost et al., 2011*) can be used to generate anchor gene pairs.

(3) dS value calculations: based on alignments of CDS, GenoDup can automatically build a control file, required by Codeml, for each gene pair. Codeml package in PAML is used to calculate dS values with parameters: noisy = 9, verbose = 1, runmode = −2, seqtype = 1, CodonFreq = 2, model = 0, NSsites = 0, icode = 0, fix_kappa = 0, kappa = 1, fix_omega = 0, and omega = 0.5.

All of the three subroutines above can automatically run in the GenoDup Pipeline, and finally generate two directories and a text file as outputs. One directory called aln_genefam contains all CDS alignments in Fasta format. The other directory, called PAML_result, contains all results generated by Codeml. A text file called dS_value.txt contains all dS values of gene pairs. An R script (plot_GenoDup.r) is also provided to plot dS distributions.

RESULTS

Empirical data validation

To evaluate the performance of the GenoDup Pipeline, I applied it to two empirical data: one is a model plant (*Arabidopsis thaliana*) and the other is a model animal (*Oncorhynchus mykiss*). *Arabidopsis thaliana* has undergone two independent WGDs (alpha and beta WGD) and *Oncorhynchus mykiss* has experienced four independent WGDs (Ss4R, Ts3R, and Two-rounds WGD) (*Berthelot et al., 2014*; *Vanneste et al., 2014*).

The CDS, protein sequences, and genome annotation files of *Arabidopsis thaliana* were downloaded from Ensembl Plants (http://plants.ensembl.or/Arabidopsis_thaliana/Info/Index). Orthogroup information was generated with OrthoMCL and 48,307 genes of *Arabidopsis thaliana* were clustered into 5,962 orthogroups. N was set as 15, meaning that the orthogroups containing less than 15 genes were used to build gene pairs, and 68,231 paralogous gene pairs were generated in total. The entire analysis ran in 8 .8 h with 4 cores. The result showed a clearly visible peak (dS value range: 0.5∼1) in the dS distribution of paralogous gene pairs (Fig. 2A). On the other hand, MCScanX was used to generate 99,309 anchor gene pairs and the entire analysis ran in 76.8 h with 4 cores (Table 1). The result showed the same peak (dS value range: 0.5∼1) in the dS distribution of anchor gene pairs (Fig. 2B). Based on assumptions of the dS-based method, the peak represents a WGD event and the Genodup Pipeline properly detected a WGD event (alpha WGD) in *Arabidopsis thaliana.*

The CDS, protein sequences, and genome annotation files of *Oncorhynchus mykiss* were downloaded from GENEOSCOPE (http://www.genoscope.cns.fr/trout/data/). Orthogroup information was generated by OrthoMCL and 46,585 genes of *Oncorhynchus mykiss* were clustered into 6,562 orthogroups. N was set as 15, meaning that the orthogroups containing less than 15 genes were used to build gene pairs, and 42,888 paralogous gene pairs were generated in total. The entire analysis ran in 7.63 h with four cores. The result showed two clearly visible peaks (dS value ranges: 0.1∼0.5 and 1.2∼2) in the dS distribution of paralogous gene pairs (Fig. 3A). On the other hand, MCScanX was used to generate 26,880

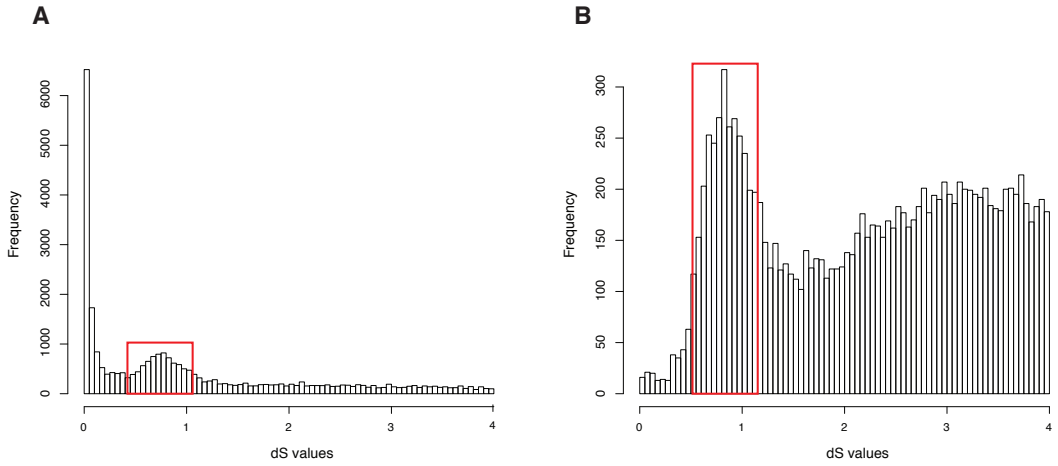

Figure 2 **dS distributions of paralogous gene pairs and anchor gene pairs in *Arabidopsis thaliana*.** (A) The peak (dS value range: 0.5∼1) marked with a red box represents a signal as alpha WGD in the dS distributions of paralogous gene pairs generated by orthogroups. (B) The peak (dS value range: 0.5∼1) marked with a red box represents a signal as alpha WGD in the dS distributions of anchor gene pairs.
Full-size DOI: 10.7717/peerj.6303/fig-2

Table 1 **Statistics of empirical data validation in GenoDup Pipeline.**

	Arabidopsis thaliana		*Oncorhynchus mykiss*	
	Orthogroup	**Anchor gene pairs**	**Orthogroup**	**Anchor gene pairs**
The number of gene pairs	68,231	99,309	42,888	26,880
Running Time (h)*	8.88	76.8	7.63	17.65

Notes.
*The running time includes the OrthoMCL running, each of run of OrthoMCL is less than 10 min.

anchor gene pairs and the entire analysis ran in 17.65 h with four cores (Table 1). The result showed a peak (dS value range: 0.1∼0.5) in the dS distribution of anchor gene pairs (Fig. 3B).

GenoDup Pipeline represented consistent result as the previous study on *Arabidopsis thaliana* and *Oncorhynchus mykiss*, respectively (*Berthelot et al., 2014*; *Vanneste et al., 2014*), indicating that GenoDup Pipeline is a reliable tool to infer WGD using the dS-based method.

Comparison GenoDup with FASTKs

The dS-based method have been applied to many studies to infer WGD (*Jiao et al., 2012*; *Vanneste et al., 2014*; *Zhao et al., 2017*) and some software has been built to make this process easier, such as DupPipe, FASTKs, and CoGe. DupPipe and CoGe are web-based methods (*Barker et al., 2010*; *Lyons et al., 2008*), while FASTKs is an open source pipeline (*McKain et al., 2016*). Hence, I compared these two open source pipelines (FASTKs and GenoDup) on three datasets: one is an example set in FASTKs (*Typha angustifolia*) and other two are example sets in this study (*Arabidopsis thaliana* and *Oncorhynchus mykiss*).

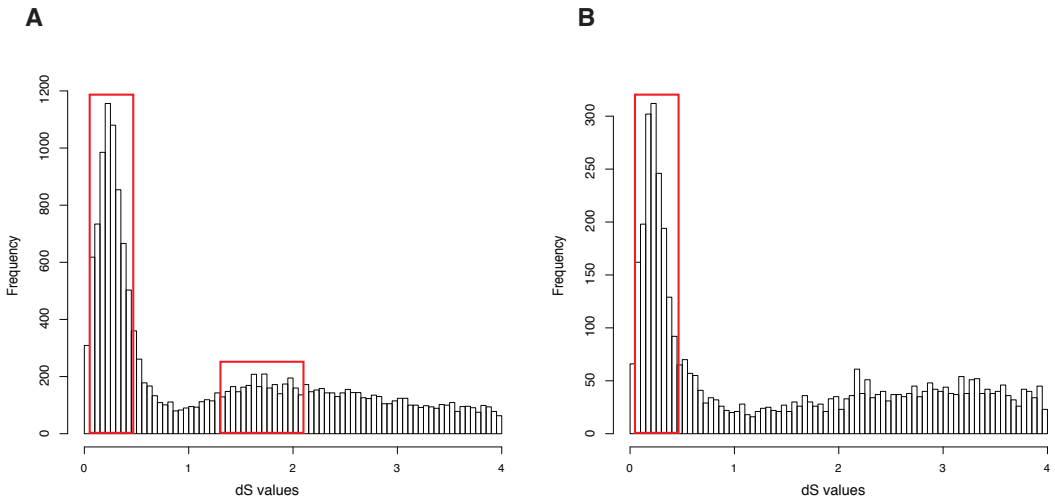

Figure 3 **dS distributions of paralogous gene pairs and anchor gene pairs in *Oncorhynchus mykiss*.** (A) The peaks (dS value ranges: 0.1∼0.5 and 1.2∼2) marked with red boxes represent signals of the Ss4R and Ts3R WGD, respectively, in the dS distributions of paralogous gene pairs generated by orthogroups. (B) The peak (dS value range: 0.1 0.5) marked with a red box represents the signal as the Ss4R WGD of Oncorhynchus mykiss in the dS distributions of anchor gene pairs.

Full-size DOI: 10.7717/peerj.6303/fig-3

The three datasets were run in FASTKs with 4 cores and default settings. While, for GenoDup Pipeline, I firstly used OrthoMCL to build the orthogroups for each dataset and run GenoDup Pipeline with four cores and N was set as 5. I found that both pipelines could infer WGD events properly (Fig. 4). In addition, I found that GenoDup Pipeline used less memory rather than FASTKs in all three datasets.

DISCUSSION

The rapid development of NGS technologies has enabled generation of massive amounts of data, allowing us to better understand the evolutionary history of all organisms. WGD is suggested to have occurred in diverse organismal groups (*Li et al., 2018*; *Van De Peer, Mizrachi & Marchal, 2017*); thus, a reliable and efficient tool to detect WGD with NGS data is greatly needed. I developed a reliable and easy-to-use tool called GenoDup Pipeline to infer WGD using the dS-based method. GenoDup Pipeline is written in Python and can be run with one command. It is easy to use for researchers who have little experience in bioinformatics.

GenoDup Pipeline runs faster when taking orthogroup information as input rather than taking gene pair information as input (Table 1). Because the total time of alignment process is usually longer for gene pairs than for orthogroups in GenoDup Pipeline. Additionally, FASTKs runs faster for *Typha angustifolia* dataset but slower in *Arabidopsis thaliana* dataset compared with GenoDup Pipeline. One possible reason is that there are many gene pairs (123,145) generated by all-against-all BLAST in *Arabidopsis thaliana* dataset than that of *Typha angustifolia* (10,119) dataset (Table 2). In other words, the running time of FASTKs

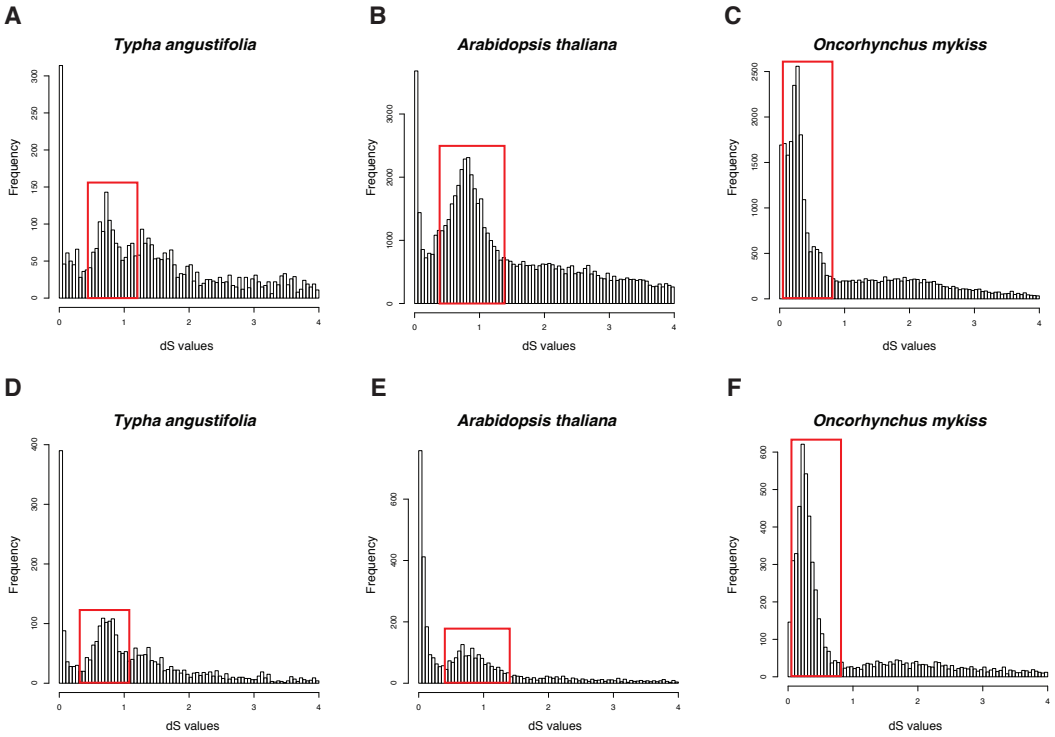

Figure 4 dS distributions of three datasets inferred by GenoDup and FASTKs. The dS distributions inferred by FASTKs on (A) *Typha angustifolia* (B) *Arabidopsis thaliana* (C) *Oncorhynchus mykiss*. The dS distributions inferred by GenoDup on (D) *Typha angustifolia* (E) *Arabidopsis thaliana* (F) *Oncorhynchus mykiss*. The peak marked with a red box represents a WGD event.

Full-size DOI: 10.7717/peerj.6303/fig-4

Table 2 Performance comparison between FASTKs and GenoDup.

	Typha angustifolia			*Arabidopsis thaliana*			*Oncorhynchus mykiss*		
	The number of gene pairs	**Running time (h)**	**Maximum memory usage (Mb)**	**The number of gene pairs**	**Running time (h)**	**Maximum memory usage (Mb)**	**The numer of gene pairs**	**Running time (h)**	**Maximum memory usage (Mb)**
FASTKs	10,119	0.4	417	123,145	9.5	1,066	43,704	1.8	1,263
GenoDup	6,853	0.8*	85	9,051	1.6*	145	8,910	1.2*	218

Notes.

*The running time includes the OrthoMCL running, each of run of OrthoMCL is less than 10 min.

is depended on the complexity of genomes. In contrast, the running time of GenoDup Pipeline is related to the parameter N (Tables 1 and 2).

The empirical validation shows that the analysis of *Arabidopsis thaliana* generated with GenoDup Pipeline presented a clearly visible signal for a WGD event (alpha WGD) but the signal of the second WGD event (beta WGD) was lost (Fig. 2). This result is consistent with previous studies because the dS methods cannot infer WGD when organisms have undergone extensive gene loss or genome shuffling, especially for plants (*Rabier, Ta & Ane, 2014*; *Tiley, Ane & Burleigh, 2016*). Additionally, the *Oncorhynchus mykiss* analysis with GenoDup showed two WGD signals (Ss4R and Ts3R) with orthogroup information

(Fig. 3A) (*Rabier, Ta & Ane, 2014*). Yet, The Ts3R signal was lost in the distribution of anchor gene pairs because there were few anchor gene pairs in the analysis (Fig. 3B, Table 1). Moreover, the analyses on *Oncorhynchus mykiss* using orthogroups and anchor gene pairs did not present the two-round WGDs because the two-round WGDs are too ancient to infer with the dS-based method. Importantly, the dS-based method is widely debated for generating artificial signals, as a result of dS saturation when the dS value >1 (*Vanneste, Van de Peer & Maere, 2012*). In addition, the dS-based method is not useful to infer WGD on genomes which have lower gene retention rates. As well, the young WGD events are not easy to be inferred because of allelic variations (*Tiley, Barker & Gordon Burleigh, 2018*). Thus, GenoDup Pipeline is suitable to infer WGD when the dS value <1. However, other evidence from phylogenetic analysis and/or synteny block analysis is also needed for drawing a conclusive result.

In all, This study presents a reliable and user-friendly tool to infer WGD using the dS-based method, beneficial for large developing sequencing projects (1KP, 10KP-EBP, 1KITE, and i5K) (*Li et al., 2018*; *Van De Peer, Mizrachi & Marchal, 2017*).

ACKNOWLEDGEMENTS

I thank Dr. Noriyuki Satoh and Dr. Evan P. Economo for their comments on the project. I thank Dr. Steven D. Aird for editing the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was supported by OIST and was funded by a JSPS grant (No. 17J00557 to Yafei Mao). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the author:
OIST.
JSPS: 17J00557.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Yafei Mao conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:
GitHub: https://github.com/MaoYafei/GenoDup-Pipeline.

REFERENCES

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* **38**:W7–13 DOI 10.1093/nar/gkq291.

Barker MS, Dlugosch KM, Dinh L, Challa RS, Kane NC, King MG, Rieseberg LH. 2010. EvoPipes. net: bioinformatic tools for ecological and evolutionary genomics. *Evolutionary Bioinformatics* **6**:EBO–S5861 DOI 10.4137/EBO.S5861.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, Aury JM. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications* **5**:3657 DOI 10.1038/ncomms4657.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422(6930)**:433–438 DOI 10.1038/nature01521.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16(1)**:157 DOI 10.1186/s13059-015-0721-2.

Glasauer SMK, Neuhauss SCF. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics* **289(6)**:1045–1060 DOI 10.1007/s00438-014-0889-2.

Jiao Y. 2018. Double the genome, double the fun: genome duplications in angiosperms. *Molecular Plant* **11(3)**:357–358 DOI 10.1016/j.molp.2018.02.009.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, Wu X. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* **13(1)**:R3 DOI 10.1186/gb-2012-13-1-r3.

Jiao Y, Li J, Tang H, Paterson AH. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell* **26**:2792–2802.

Jiao Y, Paterson AH. 2014. Polyploidy-associated genome modifications during land plant evolution. *Philosophical Transactions of the Royal Society of London. Series B* **369(1648)**:20130355 DOI 10.1098/rstb.2013.0355.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473(7345)**:97–100 DOI 10.1038/nature09916.

Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research* **30(14)**:3059–3066 DOI 10.1093/nar/gkf436.

Li L, Stoeckert Jr CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13(9)**:2178–2189 DOI 10.1101/gr.1224503.

Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods.

Proceedings of the National Academy of Sciences of the United States of America **115(18)**:4713–4718.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290(5494)**:1151–1155 DOI 10.1126/science.290.5494.1151.

Lyons E, Pedersen B, Kane J, Freeling M. 2008. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Tropical Plant Biology* **1(3–4)**:181–190 DOI 10.1007/s12042-008-9017-y.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102(15)**:5454–5459 DOI 10.1073/pnas.0501102102.

McKain MR, Tang H, McNeal JR, Ayyampalayam S, Davis JI, DePamphilis CW, Givnish TJ, Pires JC, Stevenson DW, Leebens-Mack JH. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biology and Evolution* **8(4)**:1150–1164 DOI 10.1093/gbe/evw060.

Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K. 2011. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research* **40(2)**:e11–e11 DOI 10.1093/nar/gkr955.

Rabier CE, Ta T, Ane C. 2014. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular Biology and Evolution* **31(3)**:750–762 DOI 10.1093/molbev/mst263.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, De Pamphilis CW, Wall K, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96(1)**:336–348 DOI 10.3732/ajb.0800079.

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* **35**:119–125 DOI 10.1016/j.gde.2015.11.003.

Tiley GP, Ane C, Burleigh JG. 2016. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biology and Evolution* **8(4)**:1023–1037 DOI 10.1093/gbe/evw058.

Tiley GP, Barker MS, Gordon Burleigh J. 2018. Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biology and Evolution* **10(11)**:2882–2898 DOI 10.1093/gbe/evy200.

Van De Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18(7)**:411–424 DOI 10.1038/nrg.2017.26.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous—paleogene boundary. *Genome Research* **24(8)**:1334–1347 DOI 10.1101/gr.168997.113.

Vanneste K, Van de Peer Y, Maere S. 2012. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* **30(1)**:177–190 DOI 10.1093/molbev/mss214.

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**(**7**):e49–e49 DOI 10.1093/nar/gkr1293.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(**8**):1586–1591 DOI 10.1093/molbev/msm088.

Zhao G, Zou C, Li K, Wang K, Li T, Gao L, Zhang X, Wang H, Yang Z, Liu X, Jiang W. 2017. The Aegilops tauschii genome reveals multiple impacts of transposons. *Nature Plants* **3**(**12**):946–955 DOI 10.1038/s41477-017-0067-8.